ISSN 2367-5721

SOCIOBRAINS

INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
ISSUE 4, DECEMBER 2014

PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
JULIAN VASILEV 66 - 88

NETWORK SECURITY BY ANALISIS OF LOG FILES OF APACHE WEB SERVER

JULIAN VASILEV

ASSOCIATE PROFESSOR PHD
DEPARTMENT OF INFORMATICS
VARNA UNIVERSITY OF ECONOMICS

BULGARIA
VASILEV@UE-VARNA.BG

ABSTRACT: THE PURPOSE OF THIS STUDY IS TO PROVIDE AN INNOVATIVE APPROACH OF CONDUCTING NETWORK SECURITY BY AN ANALYSIS OF THE LOG FILES OF APACHE WEB SERVER. VARIOUS METHODS AND TOOLS FOR ARE STUDIED. BY THE APPLICATION OF QUANTITATIVE, QUALITATIVE AND STATISTICAL METHODS SEVERAL HACKER ATTACKS ARE IDENTIFIED.

KEY WORDS: LOG FILE, WEB SERVER APACHE, NETWORK SECURITY, ANALYSIS OF LOG FILES, SPSS, QUANTITATIVE METHODS, QUALITATIVE METHODS

JEL: C55, C63, C87

1. INTRODUCTION

SPECIALIZED software systems are used in our everyday life. All web servers use log files. All queries to the web server are logged to one or several log files. As the time passes, the size of log files increases. That is why spreadsheets may not be used to analyze log files. That is why specialized software products are used to analyze big data stored in log files. Some companies afford themselves the creation of their own software packages for the analysis of log files.

NETWORK security is aimed at finding problem areas. Some queries of end-users may be treated as hacker attacks, others – not. A differentiation between normal and abnormal user activity has to be made. It is a difficult task. A lot of queries from one IP address may be a hacker's attack or not. For instance a lot of students use one and the same IP address when they are at the university to check their e-mail. In this case we have a lot of queries from one IP address which are not a hacker's attack.

AN analysis of log files should indicate time points when the server is significantly loaded. The analysis can show and requested access to web pages that no longer exist on the web server. It is possible to carry out an analysis in terms of usability server from separate geographic regions. The analysis can show the behavior of individual users.

THE methods and tools for analysis of log files are the main subject of this study. This research aims at finding problems with the log files of the web server Apache, installed on the server "bi.ue-varna.bg". The server has been working for five years (2009 to 2014). The collected data in the log files can be analyzed.

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

THE main objective of the study is focused at identification of problem areas in the operation of the web server Apache. Problem areas may be: weaknesses in network security, large load at peak times and inability to service multiple users. In order to solve the target for the purposes of this study we need to address the following main topics:

1. Making a study of the theoretical aspects of the Web server with respect to the generation of log files.
2. Making a survey of known methods and software tools for the analysis of log files.
3. Formulating of a methodology for analyzing log files.
4. Applying the proposed methodology for analysis of log files.

2. LITERATURE REVIEW

DIFFERENT kind of attacks may be received by a web server. The regular work of end-users of a server may be hardened by attacks of hackers. Attacks are processed by the web server as regular requests. Thus some authors (Jeyanthi, N. et. al., 2014) propose models for preventing denial of service (DoS) attack. Their model detects the attack at an earlier stage. It also helps to prevent from further attack. A kind of Recurrence Quantification Analysis (RQA) is carried out to analyze the behavior of traffic data. Flooding attacks may be prevented. The authors find that the RQA method is efficient in detecting denial of service attacks in their early stage.

NETWORKS are widely spread. A lot of people use wireless networks. Some Wi-Fi networks are open to many users. Sniffing tools may be used by hackers to track network packets. Session hijacking is one of the most common attacks. Some authors (Manivannan, S. and Sathiyamoorthy, E., 2014) proposed the strong and encrypted session ID to prevent the session hijack attacks in web applications. They proved that 212 characters encrypted session ID completely prevents the session hijack attacks in web applications of wireless networks.

CLOUD technologies are quite popular in recent years. The use of software as a service (SaaS) means that corporate data is stored in outside servers. The security issues concerning network security in cloud applications are new and very interesting. One of the most serious threats is the distributed denial of service (DDoS) attack. Two researchers (Jeyanthi, N. and Mogankumar P., 2014) propose a new mechanism to protect the cloud infrastructure from attacks.

DATA centers store a lot of data. Many users have access to the data centers. This fact shows that data centers are open to many users – real users and hackers. Some authors (Jeyanthi, N. and Iyengar, Ch., 2013) propose algorithms to detect hacker's attack by analyzing traffic. Attacks are stopped an early stage.

NETWORK security issues are widely studied. A lot of methods are created to recognize current or past threats. A group of authors (Lv H. et.al., 2013) propose a threat identification model. The threat identification model is called Attack State Transition Graph and Real-Time Attack State Graph. It is constructed by an Expanded Finite-State Automat. A new threat identification algorithm is presented. Valid threats are obtained by correlating the dynamic alarms with a static attack scenario.

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

3. Experimental analysis

3.1. An analysis of Access.log file of web server Apache

For the purpose of this study we examine the files access.log and error.log, created by the web server Apache, installed on server http://bi.ue-varna.bg:8080. The Access.log file contains 35218 rows. The Error.log file contains 20981 lines. Both files are created in the period 18 Mar 2010 – 28 Sep 2014. The bi.ue-varna.bg server is used for the installed IIS web server, Apache web server and several web applications. The focus of the current study is the two files from the web server Apache. Future research may be directed to jointly study and analyzed on log files from both the web server (Apache and IIS).

An analysis of log files can start by transferring them in the middle of a spreadsheet. The analysis begins with the file Access.log. The data is in the following format:

127.0.0.1 - - [18/Mar/2010:16:21:35 +0200] "GET /xampp/xampp.css HTTP/1.1" 200 4383 "http://localhost/xampp/splash.php" "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; WOW64; Trident/4.0;

After transferring the data from a text file into a spreadsheet, some columns are removed and the following data is left.

Table 1.

A part of the data stored in the Access.log

IP address	Data and time	Access to file	Browser	URL address
127.0.0.1	[18/Mar/2010 16:21:35	GET / HTTP/1.1	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; WOW64	-
127.0.0.1	[18/Mar/2010 16:21:35	GET /xampp/ HTTP/1.1	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; WOW64	-
127.0.0.1	[18/Mar/2010 16:21:35	GET /xampp/splash.php HTTP/1.1	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; WOW64	-
127.0.0.1	[18/Mar/2010 16:21:35	GET /xampp/xampp.css HTTP/1.1	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; WOW64	http://localhost/xampp/splash.php
127.0.0.1	[18/Mar/2010 16:21:35	GET /xampp/img/xampp-logo.jpg HTTP/1.1	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; WOW64	http://localhost/xampp/splash.php
127.0.0.1	[18/Mar/2010 16:21:35	GET /xampp/img/blank.gif HTTP/1.1	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; WOW64	http://localhost/xampp/splash.php

So far we have implemented procedures for clearing the data – extract, transform and load (ETL). When transferring data from the txt file Access.log to rows and columns within the spreadsheet the used separator symbol for columns is interval (ASCII code 32). The separator for rows is carriage return and line feed symbols (ASCII codes 10 and 13). To begin the analysis we may start with the removal of number of requests from unique IP addresses. By means of a Pivot Table we find that requests to the server are derived from 1490 unique IP addresses.

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

Table 2.

A part of a Pivot table, showing the number of queries, sent by unique IP addresses

IP address	Number of queries to the server
1.245.71.56	1
1.85.38.178	1
100.42.230.66	20
101.228.205.162	1
101.23.107.111	1
101.68.90.25	1
101.78.134.166	5

IT is noteworthy that there are remote users having legal access to the server only once or twice. This feature requires the data to be filtered. When filtering rows of the Pivot Table giving the criteria "over 10 requests to the server", the output contains 145 of 1490 lines. When filtering rows of the Pivot Table, giving the criteria "over 20 requests to the server", the output consists of 101 of 1490 lines. When filtering the rows of the Pivot Table by giving the criteria "over 30 requests to the server" 72 lines are displayed of total 1490 lines. When filtering the dataset in the Pivot Table with the criteria "over 40 requests to the server" 61 resulting rows are displayed of total 1490 lines. When filtering the rows in the Pivot Table with the criteria "over 100 requests to the server" 20 rows are displayed of total 1490 lines.

THE report may be sorted in the descending order of the number of received requests.

Table 3.

A report showing the count of received queries by unique IP addresses

IP address	count of queries
5.9.87.141	12180
67.105.202.226	3582
194.141.79.79	2875
127.0.0.1	1461
85.105.164.143	1000
87.126.36.237	669
95.224.74.85	643
62.225.239.236	639
79.100.173.232	609
211.199.115.61	584
192.168.32.1	506
5.178.86.77	382
87.126.47.183	362

192.168.32.74	350
84.43.152.114	310
62.64.104.178	290
195.20.24.1	274
95.42.118.183	231
213.91.247.172	208
213.91.163.5	179
90.154.197.55	174
79.141.160.15	156
71.86.30.194	155
195.110.42.246	145
60.173.14.85	142
94.190.193.13	141
194.141.79.1	133
95.42.141.239	115
87.126.39.17	114

THE report shows that requests from localhost (127.0.0.1) are not the most of them. Our initial expectations are not fulfilled. The analysis should continue with respect to tracking IP addresses from which requests are made. For this purpose use the site www.iplocationfinder.com.

Table 4.

A report for the received queries grouped by country and IP address

IP address	Number of queries	Country/town
5.9.87.141	12180	Germany
67.105.202.226	3582	USA
194.141.79.79	2875	Bulgaria
127.0.0.1	1461	Bulgaria
85.105.164.143	1000	Turkey, Antalya
87.126.36.237	669	Bulgaria
95.224.74.85	643	Italy, Rome
62.225.239.236	639	Germany, Torgau
79.100.173.232	609	Bulgaria, Varna
211.199.115.61	584	Korea
192.168.32.1	506	Bulgaria, Varna
5.178.86.77	382	Russia
87.126.47.183	362	Bulgaria
192.168.32.74	350	Bulgaria

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

84.43.152.114	310	Bulgaria, Varna
62.64.104.178	290	Ukraine

THE presented report shows that requests to the server come from different countries around the world. Several assumptions may be made. Most requests (12180) to the server came from the following IP address: "5.9.87.141". They are about one-third of all requests to the server. The originally transferred data table can be filtered by the selected IP address. In this case, the rules for deletion of some of the columns in the log file and filtering the rows in it are obeyed. These procedures are part of the ETL process. A similar analysis can be performed for another IP address.

AFTER filtering the data by IP address, it is established that all 12180 requests from IP address "5.9.87.141" are made on 15 Dec 2012, from 09:20 pm to 09:59 pm. Obviously **we found an attack**. More than 12000 calls to the server are made within 30 minutes. There may be a qualitative analysis on the websites to which it is attempted to be opened. Obviously requests are sent by a computer program, a list of URL addresses previously been set

Table 5.
A part of the web pages, tried to be opened by the hacker, concerning the administration of the server

GET /adm/ HTTP/1.1
GET /~admin/ HTTP/1.1
GET /admin/ HTTP/1.1
GET /admin-bak/ HTTP/1.1
GET /admin-old/ HTTP/1.1
GET /admin.back/ HTTP/1.1
GET /admin_/ HTTP/1.1
GET /administration/ HTTP/1.1
GET /administrator/ HTTP/1.1
GET /adminuser/ HTTP/1.1
GET /adminweb/ HTTP/1.1
GET /authadmin/ HTTP/1.1

Table 6.
Web pages connected with security

GET /private/ HTTP/1.1
GET /protected/ HTTP/1.1
GET /secret/ HTTP/1.1
GET /secure/ HTTP/1.1
GET /secured/ HTTP/1.1
GET /siteadmin/ HTTP/1.1
GET /sites/ HTTP/1.1
GET /ssi/ HTTP/1.1
GET /ssl/ HTTP/1.1
GET /sslkeys/ HTTP/1.1

Table 7.
Web pages, connected with statistics

GET /ustats/ HTTP/1.1
GET /web_usage/ HTTP/1.1
GET /webaccess/ HTTP/1.1
GET /webadmin/ HTTP/1.1
GET /webalizer/ HTTP/1.1
GET /webalizer/NonExistent.html HTTP/1.1
GET /webstat/ HTTP/1.1
GET /webstats/ HTTP/1.1
GET /webtrends/ HTTP/1.1
GET /wstats/ HTTP/1.1

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

GET /wusage/ HTTP/1.1
GET /wwwlog/ HTTP/1.1
GET /wwwstat/ HTTP/1.1
GET /wwwstats/ HTTP/1.1
GET /~stats/ HTTP/1.1
GET /~webstats/ HTTP/1.1

Table 8.

Web pages, concerning the hypothesis of the hacker, that an e-shop is hosted on the server

GET /inventory/ HTTP/1.1
GET /grocery/ HTTP/1.1
GET /payments/ HTTP/1.1
GET /misc/ HTTP/1.1
GET /help/ HTTP/1.1
GET /order/ HTTP/1.1
GET /orders/ HTTP/1.1
GET /purchase/ HTTP/1.1
GET /purchases/ HTTP/1.1
GET /sales/ HTTP/1.1
GET /faq/ HTTP/1.1

A brute force attack is found.

The second critical IP address (67.105.202.226) has made 3582 requests of total 35218 requests. Requests are concentrated in two days – 29 Apr 2010 and 0 May 2010. The received requests are related to the verification of the installed version of PHP.

Table 9.

A check of the installed version of PHP

GET /phpMyAdmin-2.2.3/main.php HTTP/1.0
GET /phpMyAdmin-2.2.6/main.php HTTP/1.0
GET /phpMyAdmin-2.5.1/main.php HTTP/1.0
GET /phpMyAdmin-2.5.4/main.php HTTP/1.0
GET /phpMyAdmin-2.5.5-rc1/main.php HTTP/1.0
GET /phpMyAdmin-2.5.5-rc2/main.php HTTP/1.0
GET /phpMyAdmin-2.5.5/main.php HTTP/1.0
GET /phpMyAdmin-2.5.5-pl1/main.php HTTP/1.0
GET /phpMyAdmin-2.5.6-rc1/main.php HTTP/1.0
GET /phpMyAdmin-2.5.6-rc2/main.php HTTP/1.0
GET /phpMyAdmin-2.5.6/main.php HTTP/1.0
GET /phpMyAdmin-2.5.7/main.php HTTP/1.0
GET /phpMyAdmin-2.5.7-pl1/main.php HTTP/1.0
GET /phpMyAdmin-2.6.0-alpha/main.php HTTP/1.0
GET /phpMyAdmin-2.6.0-alpha2/main.php HTTP/1.0
GET /phpMyAdmin-2.6.0-beta1/main.php HTTP/1.0
GET /phpMyAdmin-2.6.0-beta2/main.php HTTP/1.0
GET /phpMyAdmin-2.6.0-rc1/main.php HTTP/1.0

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

GET /phpMyAdmin-2.6.0-rc2/main.php HTTP/1.0
GET /phpMyAdmin-2.6.0-rc3/main.php HTTP/1.0
GET /phpMyAdmin-2.6.0/main.php HTTP/1.0
GET /phpMyAdmin-2.6.0-pl1/main.php HTTP/1.0
GET /phpMyAdmin-2.6.0-pl2/main.php HTTP/1.0
GET /phpMyAdmin-2.6.0-pl3/main.php HTTP/1.0
GET /phpMyAdmin-2.6.1-rc1/main.php HTTP/1.0
GET /phpMyAdmin-2.6.1-rc2/main.php HTTP/1.0
GET /phpMyAdmin-2.6.1/main.php HTTP/1.0

THE received 2875 requests from IP address 194.141.79.79 are combined with uploading and testing site http://bi.ue-varna.bg/bi/. Such activity can be considered a hacker attack, but it is simply testing a web application. 1000 requests have been received by address 85.105.164.143. Requests are sent shortly after midnight within 10 minutes. All requests show the same experience to open folder "\ x03" server. Such activity seems to be a hacker attack, but it is not quite normal. Perhaps the response time of the server is tested. The received 669 requests from address 87.126.36.237 represent the normal activity of the end user to the server.

WE may make the assumptions that less than 500 requests from one and the same IP address are unlikely to contain a hacker attack. Assumptions may be demonstrated, proved or rejected. For example 643 requests are received from IP address "87.126.36.237 643".

Table 10.

A part of the requests from IP address 87.126.36.237

[21/Mar/2010:22:17:26	GET /db/read_dump.phpmain.php HTTP/1.0
[21/Mar/2010:22:17:26	GET /dbadmin/read_dump.phpmain.php HTTP/1.0
[21/Mar/2010:22:17:26	GET /web/phpMyAdmin/read_dump.phpmain.php HTTP/1.0
[21/Mar/2010:22:17:27	GET /admin/pma/read_dump.phpmain.php HTTP/1.0
[21/Mar/2010:22:17:27	GET /admin/phpmyadmin/read_dump.phpmain.php HTTP/1.0
[21/Mar/2010:22:17:27	GET /phpmyadmin2/read_dump.phpmain.php HTTP/1.0
[21/Mar/2010:22:17:27	GET /phpmyadmin1/read_dump.phpmain.php HTTP/1.0
[21/Mar/2010:22:17:28	GET /phpadmin/read_dump.phpmain.php HTTP/1.0
[21/Mar/2010:22:17:28	GET /myadmin/read_dump.phpmain.php HTTP/1.0
[21/Mar/2010:22:17:28	GET /phpMyAdmin-2.2.3/read_dump.phpmain.php HTTP/1.0

HERE obviously requests are sent back through a program. Requests are sent during the nighttime. It is possible to make the assumption that hackers are more active at night. There are more requests in certain days of the week than in others. Date and time in the log file is maintained in the following format "[18 / Mar / 2010: 16: 21: 35". In order to extract only the date, the MID function may be used *=MID (B2; 2; 11)* in the middle of MS Excel.

A second Pivot table is done, showing the count of requests by days.

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

Table 11.

A part of the report "Count of requests by days"

Date	Count of requests
01/Apr/2010	2
01/Apr/2013	17
01/Aug/2012	6
01/Aug/2013	4
01/Dec/2012	1
01/Dec/2013	13
01/Feb/2013	10
01/Feb/2014	3
01/Jan/2014	16
01/Jul/2012	1
01/Jul/2013	5
01/Jun/2010	61
01/Jun/2012	54
01/May/2010	736
01/Nov/2013	6
01/Oct/2012	3
01/Oct/2013	7
01/Sep/2012	2
01/Sep/2013	12
02/Apr/2010	48

THE report shows that there are days with one or two queries and days with many queries (requests to the server). The made report has 629output lines. During the 629 days of 1429 days (between 18.3.2010 and 14.9.2014) the server received queries. When filtering data by giving the criteria "over 100 requests per day", the resulting dataset contains 54 rows (relevant to 54 days). When filtering data by giving the criteria "over 10 requests a day", there are 195 rows in the resulting dataset.

THE dates in table 11 are not sorted. They are mixed because MS Excel interprets "01 / Jan / 2014" as text, not as a date. For this purpose, by replacing function (find and replace), the string "/ Jan /" is replaced by "01.". Similar adjustments are made for the remaining months. The marked transformation aims converting dates from a string into a variable of type "date".

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

Table 12.
A part of the report, containing the count of queries by days after converting the dates from "string" type to "date" type

Date	Count of requests
18.3.2010	328
19.3.2010	290
20.3.2010	35
21.3.2010	647
22.3.2010	36
23.3.2010	1227
24.3.2010	329
25.3.2010	304
26.3.2010	359
27.3.2010	18
28.3.2010	181
29.3.2010	88
30.3.2010	134
1.4.2010	2

We have deleted the records for the date 15.12.2012. On that date 12179 requests are received. We delete part of the data in the dataset to stabilize the dispersion. Currently, the count of requests per day varies between 1 and 1227. We may now display a graph of the server load (Fig. 1).

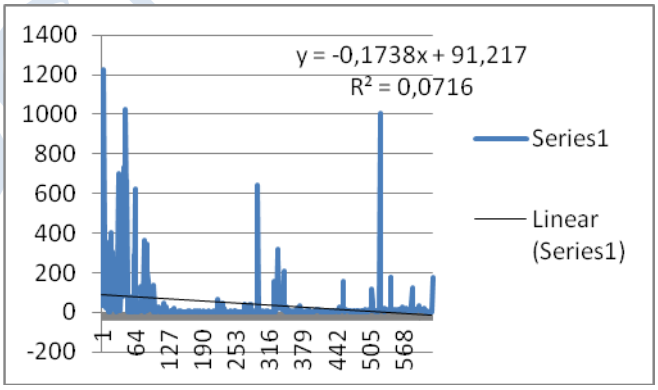

Fig. 1. *A graph, showing the server load by days*

The chart shows the server load by days. It is clear that there are days with many requests and days with little requests. The trend line shows that the total count of requests to the server decreases over time. The displayed formula (y = -0.1738x + 91.217) shows that the count of requests to the server on a daily basis decreases (-0.1738 is a negative number). The coefficient of determination R^2 has a value of 0.0716, which means that we are 7% sure of the

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

trend line. This fact can be interpreted in another way. In 93% of cases, the count of requests to the server (calculated on a daily basis) is influenced by other factors – not only of the time.

To make a second check for the presence of trend (whether the count of requests to the server is affected by time) we transfer the data shown in Table 12 from Excel in SPSS. In SPSS we create two variables – data and broj_zayavki. The first variable type is "date" and the second – an "integer" data type. The second variable (broj_zayavki) stores information about the count of requests on a daily basis. In order to further stabilize the dispersion we use the logarithmic function (Transform / Compute variable). Use the following formula

=ln(broj_zayavki)

The new variable is named ln_br_zayavki.

We check if the logarithmic value of the count of requests per day (on a daily basis) is influenced by time (Analyze/Regression/Curve estimation).

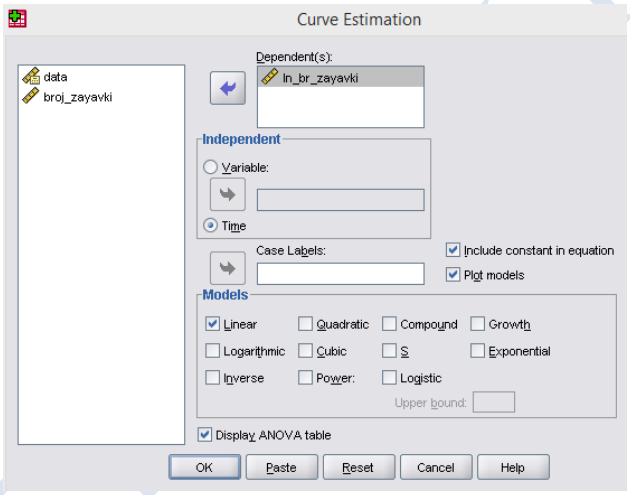

Fig. 2. *A check for tendency in SPSS*

The result of the execution of the check is visualized in the following tables.

Table 13.
Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
,148	,022	,020	1,516

Table 14.
ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	31,805	1	31,805	13,844	,000
Residual	1428,978	622	2,297		
Total	1460,783	623			

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

Table 15.
Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Case Sequence	-,001	,000	-,148	-3,721	,000
(Constant)	2,346	,122		19,311	,000

THE analysis of the three tables indicates the following. The model is adequate (Sig. 000 <0.05). R^2 has a value of 0.022, which means that besides the time there are other factors affecting the count of queries per day. Coefficients in the equation are statistically significant. The value of the coefficient B has a negative value, which could be interpreted that the number of requests to the server decreases when time passes (Fig. 3).

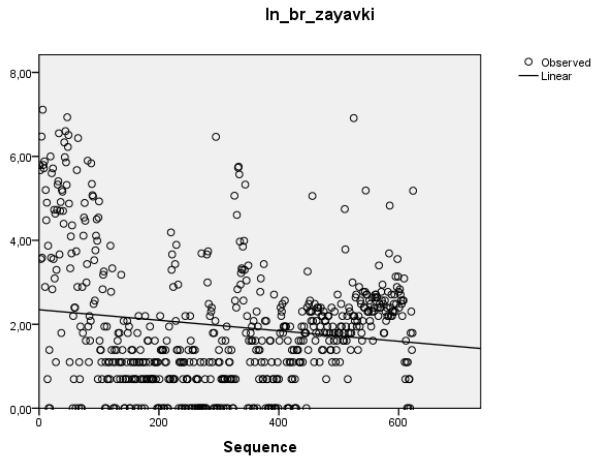

Fig. 3. *Distribution of the count of queries by days*

THE distribution of requests throughout the days shows that requests to the server are completely random in nature. Clustered dots do not stand in any part of the scatter diagram.

Similar checks have to be performed excluding the constant in the equation (Fig. 4).

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
ISSUE 4, DECEMBER 2014
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
JULIAN VASILEV 66 - 88

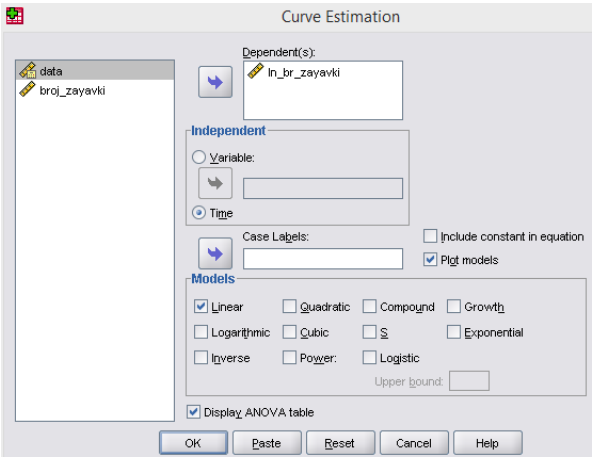

Fig. 4. *A check for tendency without including the constant in the equation*

Table 16.

Model summary excluding the constant in the equation

Model Summary[a]			
R	R Square	Adjusted R Square	Std. Error of the Estimate
,637	,406	,405	1,915

a. The equation was estimated without the constant term.

Table 17.

ANOVA table showing information for the model excluding the constant from the equation

ANOVA[a]					
	Sum of Squares	df	Mean Square	F	Sig.
Regression	1559,061	1	1559,061	424,947	,000
Residual	2285,684	623	3,669		
Total	3844,745	624			

a. The equation was estimated without the constant term.

Table 18.

Model coefficients without including the constant in the equation

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Case Sequence	,004	,000	,637	20,614	,000

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

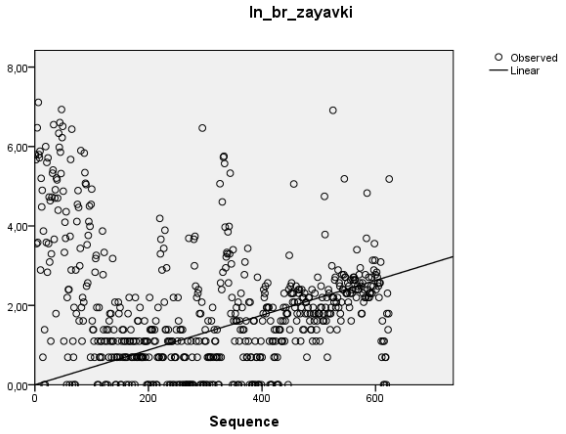

Fig. 5. *Distribution of count of queries to the server by days, without including the constant in the equation*

When excluding the constant of the equation, another model is prepared. It is statistically significant. R^2 has a higher value (0.406). The new model still shows that in addition to the time there are other influencing factors on the count of requests that are received by the server.

We may assume that "month" and "year" are possible factors affecting the count of queries. By testing statistical hypotheses, we can check if the month and year affect the number of requests received by the server. Data grouped by days (624 lines) may be grouped further by year and month by a third Pivot table.

Table 19.
Distribution of the server load by years and months

Years and months	Count of queries (requests to the server)
2010	**13713**
3	3976
4	4426
5	4880
6	431
2012	**2695**
5	1515
6	504
7	98
8	83
9	70
10	54
11	306
12	65
2013	**5859**

1	185
2	715
3	329
4	1503
5	142
6	62
7	148
8	357
9	214
10	352
11	1483
12	369
2014	**770**
1	553
2	217

THE data presented in the Pivot Table should be converted to be able to be analysed in SPSS.

Table 20.

Converted data by months and years

Year	Month	Count of requests
2010	3	3976
2010	4	4426
2010	5	4880
2010	6	431
2012	7	1515
2012	8	504
2012	9	98
2012	10	83
2012	11	70
2012	12	54
2012	13	306
2012	14	65
2013	1	185
2013	2	715
2013	3	329
2013	4	1503
2013	5	142
2013	6	62
2013	7	148
2013	8	357

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

2013	9	214
2013	10	352
2013	11	1483
2013	12	369
2014	1	553
2014	2	217

THE transformed data may be transferred to SPSS to make further analysis. The analysis includes verification of single-factor models: (1) to check if the year affects the number of received requests and (2) to check whether the month has regular influence on the count of queries per month. Then the analysis may continue with two-factor models and calculation of seasonal fluctuations.

IN order to eliminate seasonality, first a time series is defined (Data/Define dates). The natural logarithm is calculated (to stabilize the dispersion in the Series). Seasonal fluctuations are calculated (Analyze/Forecasting/Seasonal Decomposition).

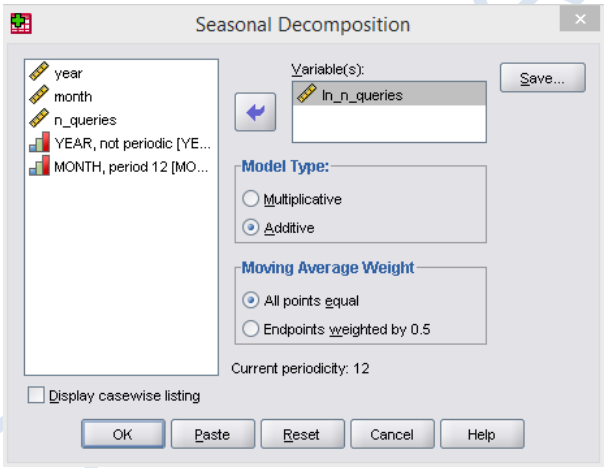

Fig. 6. *Calculation of seasonal fluctuations*

THE additive model is selected. Because we do not have data for the year 2011 (probably then the server has not received requests) and for the entire 2014 we cannot calculate the index of seasonality. For the calculation, it takes at least four full years. It is necessary to check for autocorrelation.

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

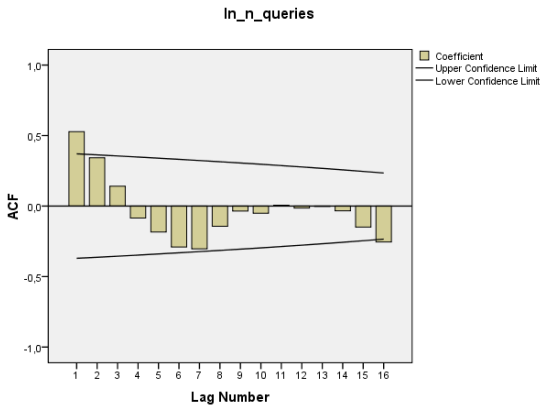

Fig. 7. *A check for autocorrelation*

SOME coefficients intersect the upper and lower limit of the confidence interval. It means that there is autocorrelation in the series. We have to calculate the first differences to remove the autocorrelation (Transform/Create time series). We make again a check for autocorrelation.

THE time series containing the first consecutive differences may be used to perform a regression analysis. First, the regression analysis is carried out by turning on the constant in the equation. The ANOVA table shows that the model is not statistically significant (Sig. 0.341). Second, the regression analysis is carried out without involvement of the constant in the equation. The ANOVA table shows that the model is not adequate (Sig. 0.638). We may conclude that **the year has no impact on the count of requests received by the server**.

THE analysis continues with checking the assumption that the month has an influence on the count of requests received by the server (Analyze/Regression/Linear). First, the regression analysis is carried out by including the constant in the equation. The ANOVA table shows that the model is not statistically significant (Sig. 0.634). Second, the regression analysis is carried out without including of the constant in the equation. The ANOVA table shows that the model is not statistically significant (Sig. 0.520). We conclude that **the month has no impact on the count of requests received by the server**.

THE following conclusions may be made, based on the study in 3.1:

1. To make the analysis of data from log files, the data should be converted and filtered. Extract, transform, load (ETL) procedures have to be carried out.

2. The requests to the server are from different countries. Most of the requests are made by computer programs of hackers in the nighttime.

3. MS Excel and SPSS may be used as appropriate instruments for analyzing the log files.

4. An analysis of the log file includes application of quantitative methods and qualitative methods.

5. One-third of all requests are made by a hacker at night just for 30 minutes.

6. One of the most common methods of hackers is "brute force".

7. We cannot detect a relationship between time and total number of requests received by the server. Time should not be used as an independent variable in a model with dependent

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

variable "count of queries". It is proved that the number of requests to the server neither increases nor decreases throughout the time. There are great variations in the count of requests, calculated on a daily basis.

8. The year and the month (in which the request is made to the server) do not have a logical impact on the activity of end-users of the server and hackers (who attack the server).

3.2. Analysis of the Error.log file of Apache web server

The analysis of the error.log files my show bugs in software applications as well as hacker attacks. There a lot of types of software errors, that may be recorded in the error.log file. For instance if an end-user tries to open a removed web site from a corporate web site, the error.log file contains one row to the wrong reference. The row contains the text „File does not exist".

[Wed Feb 12 20:19:03 2014] [error] [client 95.78.167.28] File does not exist: C:/xampp/htdocs/admin

There are many formats of the error.log file. The example above shows the date and time of the received request, the end-user IP address, and the file or the folder he/she tried to open.

If we look in the file (its next rows) we see that there are references to non-existing web pages.

[Thu Mar 18 16:36:27 2010] [error] [client 192.168.32.74] File does not exist: C:/xampp/htdocs/bi/highlight.js, referer: http://192.168.32.74/bi/cms_index.php
[Thu Mar 18 16:37:26 2010] [error] [client 192.168.32.74] File does not exist: C:/xampp/htdocs/bi/highlight.js, referer: http://192.168.32.74/bi/cms_add_act.php
[Thu Mar 18 16:37:26 2010] [error] [client 192.168.32.74] File does not exist: C:/xampp/htdocs/bi/js/ckeditor.js, referer: http://192.168.32.74/bi/cms_add_act.php
[Thu Mar 18 16:37:42 2010] [error] [client 192.168.32.74] File does not exist: C:/xampp/htdocs/bi/highlight.js, referer: http://192.168.32.74/bi/cms_edit_profile.php

All requests are sent by the developer of the site http://bi.ue-varna.bg/bi in testing the web site and its hyperlinks. Such a situation of testing and tracing a website or a web application is normal to generate rows in the error.log file.

20980 rows are stored in the Error.log file. To verify if there is a hacker attack, we may create a Pivot table which shows the unique IP addresses from which requests were made to the server. Unique IP addresses are 819. Upon filtering the data (IP addresses, which are derived from more than 20 queries to the server), the output contains 41 IP address. Data can be sorted in descending order of received requests. We may check the country from which each request is received.

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
ISSUE 4, DECEMBER 2014
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
JULIAN VASILEV 66 - 88

Table 21.

Count of requests, grouped by IP address, recorded in the error.log file

IP адрес	Count of rows in the error.log file	Country/Town
5.9.87.141	9446	Germany
67.105.202.226	3582	USA
194.141.79.79	723	Bulgaria
95.224.74.85	643	Italy, Rome
211.199.115.61	584	Korea
5.178.86.77	382	Russia
62.225.239.236	316	Germany, Torgau
62.64.104.178	289	Ukraine
87.126.36.237	155	Bulgaria
79.141.160.15	132	Poland, Warsaw
195.110.42.246	122	Spain
95.42.118.183	122	Bulgaria, Varna
79.100.173.232	113	Bulgaria, Varna

THE data show that the registered requests stored in the error.log file come from around the world. Table 21 shows only the IP addresses that have made more than 100 lines in the error.log file. Again, we may make the assumption that the requests to the server (recorded in the error.log file) in the nighttime are more than those made during the day. We have to mark that each request to the web server Apache generates a row in the access.log file. Some of the requests generate rows in the error.log file. Bugs in web applications (installed on the same server) listening on different ports than the web server are not recorded in the error.log file. The end-user activity of these web applications is not stored in the access.log file. Thus, these web applications should have their own log files to record end-user access and errors.

Table 22.

Count of rows in the error.log file, grouped by the time of the day

Time (hour)	Count of rows
0	115
1	89
2	43
3	101
4	730
5	82
6	60
7	85
8	1110
9	182

10	300
11	419
12	574
13	1364
14	480
15	427
16	869
17	355
18	941
19	485
20	306
21	9675
22	785
23	510

MOST errors (9675) are recorded between 21:00 and 22:00. Immediately after this critical time the range is from 13:00 to 14:00 (1364 errors) and from 08:00 to 09:00 (110 errors). The errors that are produced during the day are in most cases due to the testing of web sites. Errors recorded at night are mostly due to hacker attacks. The IP address 5.9.87.141 generated the most errors in the error.log file.

Table 23.

A part of the non-existing web pages, tried to be opened by an anonymous hacker

C:/xampp/htdocs/cgi-bin2
C:/xampp/htdocs/cgi-csc
C:/xampp/htdocs/cgi-lib
C:/xampp/htdocs/cgi-local
C:/xampp/htdocs/cgi-scripts
C:/xampp/htdocs/cgi-shl
C:/xampp/htdocs/cgi-shop
C:/xampp/htdocs/cgi-sys
C:/xampp/htdocs/cgi-weddico
C:/xampp/htdocs/cgi-win
C:/xampp/htdocs/cgibin
C:/xampp/htdocs/cgilib
C:/xampp/htdocs/cgis
C:/xampp/htdocs/cgiscripts
C:/xampp/htdocs/cgiwin

WE meet again with a "brute force" attack. 5 requests per second are sent to the server. All of them are recorded in the error.log file. It certainly means that the hacker used a computer program to send requests to the server to extract unauthorized information from it.

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
ISSUE 4, DECEMBER 2014
JULIAN VASILEV 66 - 88

THE qualitative analysis of the data in the error.log file shows that the most common error is "File does not exist". Beside it in the error.log file we find another error "The given path contained wildcard characters. Access to * failed". Single experiments were aimed at scripting ("Script not found or unable to start"). Some requests contain opinions without name of sender ("Client sent HTTP / 1.1 request without host name"). Another type of error is ("Directory index forbidden by options directive").

THE requests from IP address 67.105.202.226 also contain a brute-force attack.

Table 24.

A small part of the errors, registered from IP address 67.105.202.226

C:/xampp/htdocs/db
C:/xampp/htdocs/web
C:/xampp/htdocs/PMA
C:/xampp/htdocs/admin
C:/xampp/htdocs/dbadmin
C:/xampp/htdocs/PMA2006
C:/xampp/htdocs/pma2006
C:/xampp/htdocs/sqlmanager
C:/xampp/htdocs/mysqlmanager
C:/xampp/htdocs/p
C:/xampp/htdocs/PMA2005
C:/xampp/htdocs/pma2005
C:/xampp/htdocs/phpmanager
C:/xampp/htdocs/php-myadmin
C:/xampp/htdocs/phpmy-admin
C:/xampp/htdocs/mysql
C:/xampp/htdocs/myadmin
C:/xampp/htdocs/webadmin
C:/xampp/htdocs/sqlweb
C:/xampp/htdocs/websql
C:/xampp/htdocs/webdb
C:/xampp/htdocs/mysqladmin

REQUESTS from IP address 95.224.74.85 (which is in Italy is the fourth highest count of errors in the error.log file) largely resemble requests received from IP address 67.105.202.226, which is in the US (and the second highest number of errors recorded in the error.log file). Here we assume that hackers located in different parts of the world use the same "dictionary" to attack servers running Apache web server. It is possible that the same hacker have Trojan horses (a kind of computer viruses) on servers in different countries and from there they send requests to the observed server.

THE queries received from Korea are from IP address 211.199.115.61. Consecutive requests (60 queries) are recorded in just one second to the folder "C:/xampp/htdocs/administrator". Requests are received between 16:20 and 16:21 on 21 Mar 2010. Similarly there are sent consistently between 100 and 200 requests to the folders

ISSN 2367-5721
SOCIOBRAINS
INTERNATIONAL SCIENTIFIC ONLINE JOURNAL
WWW.SOCIOBRAINS.COM
ISSUE 4, DECEMBER 2014
PUBLISHER:
"SMART IDEAS – WISE DECISIONS" Ltd., BULGARIA
JULIAN VASILEV 66 - 88

"C:/xampp/htdocs/db", "C:/xampp/htdocs/database", "C:/xampp/htdocs/mysql" and "C:/xampp/ htdocs/admin". A similar type of hacking behaviour appears largely logical, although it is not quite traditional. Maybe hackers are expected to display the folder they are looking for in a certain time period that coincides with the time of the attack.

IT is interesting to trace the behavior of a Russian hacker (IP address 5.178.86.77). He (She) alternated attempts to open a file ("C:/xampp/htdocs/azenv.php") with queries that do not contain the name of the sender ("Sent HTTP/1.1 request without hostname"). Here we assume that the attacker has launched two programs that attack the observed server in the same time period.

THE requests received by a German hacker (IP address 62.225.239.236) contain a request to open the address (http://194.141.79.11:8080/webdav/webdav-help.php), which need authentication ("Client used wrong authentication scheme"). The 316 requests were made between 12:55 and 12:59 h. For four minutes (240 seconds) 316 lines are registered in the error.log file. It is a clear indicator that the requests are sent by a hacker through a program.

4. CONCLUSIONS

THE analysis of log files needs to be conducted to verify the network security. The necessity to analyze log files is proved. In theoretical terms a number of methods and software tools for the analysis of log files are developed. This paper presents a methodology for analyzing log files access.log and error.log web server Apache. The proposed methodology consists of qualitative as well as quantitative methods. Both quantitative analysis methods and qualitative methods of analysis are applied. Because of the relatively small volume of the two log files, the MS Excel spreadsheet and the statistical software SPSS are used.

THE existence of several hacker attacks is found as a result of the analysis. The most common method of attack is "brute force". It was found by quantitative analysis that within 30 minutes there were made 12 000 calls to the web server. Hacker attacks are committed by offenders from different countries. In most cases, they use "dictionary" of words and ready software program that send requests. Hackers located in different parts of the planet use the same "dictionary" to attack servers running Apache Web server. Sometimes hackers use two software programs to send multiple requests to a server. Qualitative analysis of the data in the file error.log shows that the most common mistake is "File does not exist."

BY statistical hypothesis testing it is found out that there is no logical relationship between time and total number of requests sent to the server. Over time, it can be argued that the number of requests to the server neither increases or decreases.

BY ANOVA we found out that the year and month (in which the request is made to the server) do not have a logical impact on the activity of users on the server and hackers who attack it.

FUTURE research can be directed to jointly study and analysis of log files from several web servers (for example Apache and IIS). Using data from the log file on a web server with other log files (showing CPU, network card and hard disk drive) may allow researchers to detect dependencies between abnormal user activity and high load of hardware resources.

References

1. **Hadzhiev, V. and others., 2009:** Statistical and econometric software. Varna: Nauka I ikonomika, 2009 (in Bulgarian).
2. **Jeyanthi, N. et. al., 2014:** RQA Based Approach to Detect and Prevent DDoS Attacks in VoIP Networks. CYBERNETICS AND INFORMATION TECHNOLOGIES, 2014, Volume 14, No 1, pp. 11-24.
3. **Manivannan, S. and Sathiyamoorthy, E., 2014:** A Prevention Model for Session Hijack Attacks in Wireless Networks Using Strong and Encrypted Session ID. CYBERNETICS AND INFORMATION TECHNOLOGIES, 2014, Volume 14, No 3, pp. 46-60.
4. **Jeyanthi, N. and Mogankumar, P. 2014:**A Virtual Firewall Mechanism Using Army Nodes to Protect Cloud Infrastructure from DDoS Attacks. CYBERNETICS AND INFORMATION TECHNOLOGIES, 2014, Volume 14, No 3, pp. 71-85.
5. **Jeyanthi, N. and Iyengar, Ch., 2013:** Escape-on-Sight: An Efficient and Scalable Mechanism for Escaping DDoS Attacks in Cloud Computing Environment. CYBERNETICS AND INFORMATION TECHNOLOGIES, 2013, Volume 13, No 1, pp. 46-60.
6. **Lv, H. et al.,2013:** Network Threat Identification and Analysis Based on a State Transition Graph. CYBERNETICS AND INFORMATION TECHNOLOGIES, 2013, Volume 13, Special issue, pp. 51-61.
7. www.iplocationfinder.com <19.11.2014>.